BỘ GIÁO DỤC VÀ ĐÀO TẠO

ĐỀ CƯƠNG LUẬN VĂN THẠC SỸ

Chuyên ngành: CÔNG NGHỆ THÔNG TIN

Mã ngành: 8480201

Đề tài:

NGHIÊN CỨU KẾT HỢP MÔ HÌNH HỌC SÂU TRANSFORMER VÀ PHÂN TÍCH TĂNG CƯỜNG ĐẶC TRƯNG DỮ LIỆU VỚI TÔ-PÔ ĐỂ NÂNG CAO HIỆU SUẤT PHÂN LỚP VĂN BẢN TIẾNG VIỆT

GVHD : TS. Phạm Thế Anh Phú

HVTH : Lưu Cang Kim Long

MSHV : 2341863008

Lớp : 23SCT31

TP. HCM, tháng 10/2025

NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

Tp. Hồ Chí Minh, ngày …… tháng …… năm 20…

Giảng viên hướng dẫn

NHẬN XÉT CỦA HỘI ĐỒNG XÉT DUYỆT

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

Tp. Hồ Chí Minh, ngày …… tháng …… năm 20…

Hội đồng xét duyệt

MỤC LỤC

I. GIỚI THIỆU 1
1.1 Đặt vấn đề 1
1.2 Các thách thức còn tồn tại 2
1.3 Lý do chọn đề tài 2
II. MỤC TIÊU, NỘI DUNG VÀ PHƯƠNG PHÁP 3
2.1 Mục tiêu của đề tài 3
2.2 Nội dung nghiên cứu 4
2.3 Các phương pháp nghiên cứu 5
III. TỔNG QUAN VỀ LĨNH VỰC NGHIÊN CỨU 7
3.1 Mô tả tập dữ liệu 7
3.2 Nghiên cứu trước đó 8
3.2.1 PhoBERT và các mô hình học sâu cho tiếng Việt 8
3.2.2 Phân tích dữ liệu tô-pô (TDA) trong NLP 8
3.2.3 Kỹ thuật tăng cường dữ liệu trong NLP 8
3.2.4 Kết hợp TDA và học sâu 9
3.2.5 Các mô hình LSTM, Bi-LSTM, GRU 9
3.3 Thiết lập và xây dựng mô hình 10
3.4 Đánh giá 11
3.5 So sánh với các phương pháp khác 12
IV. TIẾN ĐỘ THỰC HIỆN ĐỀ TÀI 12
4.1 Các bước thực hiện, thời gian bắt đầu và hoàn thành 12
4.2 Bố cục dự kiến của luận văn 13
V. TÀI LIỆU THAM KHẢO 14

DANH MỤC CÁC BẢNG

Bảng 1: Tiến độ thực hiện đề tài 13

DANH MỤC CÁC HÌNH

Hình 1:Minh họa quy trình tích hợp PhoBERT, TDA, và tăng cường dữ liệu.....10

DANH MỤC VIẾT TẮT

NLP: Natural Language Processing (Xử lý ngôn ngữ tự nhiên)

TDA: Topological Data Analysis (Phân tích dữ liệu tô-pô)

EDA: Easy Data Augmentation (Tăng cường dữ liệu dễ dàng)

PhoBERT: Pre-trained Language Model for Vietnamese (Mô hình ngôn ngữ tiền huấn luyện cho tiếng Việt)

LSTM: Long Short-Term Memory

GRU: Gated Recurrent Unit

I. GIỚI THIỆU

1.1 Đặt vấn đề

Phân lớp văn bản là một nhiệm vụ cốt lõi trong xử lý ngôn ngữ tự nhiên (NLP), với các ứng dụng quan trọng như phân tích cảm xúc, phát hiện tin giả, và tổ chức tài liệu tự động. Trong bối cảnh tiếng Việt – một ngôn ngữ ít tài nguyên với cấu trúc từ ghép, thanh điệu, và ngữ cảnh văn hóa phức tạp – bài toán phân lớp văn bản đối mặt với nhiều thách thức, bao gồm dữ liệu huấn luyện hạn chế, phân bố lớp không đồng đều, và khó khăn trong việc nắm bắt các mẫu ngữ nghĩa phức tạp. Vấn đề nghiên cứu chính của đề tài này là: Làm thế nào để nâng cao hiệu suất phân lớp văn bản tiếng Việt bằng cách kết hợp mô hình học sâu Transformer (PhoBERT) với phân tích dữ liệu tô-pô (TDA) và kỹ thuật tăng cường dữ liệu?

Các mô hình học sâu truyền thống như LSTM [1] hiệu quả trong việc xử lý dữ liệu chuỗi, nhưng bị hạn chế bởi việc chỉ tập trung vào ngữ nghĩa cục bộ và ngữ cảnh tuyến tính, bỏ qua các mối quan hệ toàn cục trong dữ liệu văn bản. Ngay cả PhoBERT [2], một mô hình Transformer tối ưu hóa cho tiếng Việt với tokenization cấp âm tiết, cũng gặp khó khăn trong các tình huống dữ liệu khan hiếm hoặc không đồng đều (ví dụ: lớp trung tính trong tập UIT-VSFC [3] chỉ chiếm ~4.32% tổng số mẫu, khoảng 690 mẫu). Những hạn chế này dẫn đến hiện tượng over-fitting và giảm khả năng phát hiện các mẫu ngữ nghĩa phức tạp, chẳng hạn như sự không nhất quán trong văn bản giả mạo hoặc các mối quan hệ giữa các đoạn văn trong phân tích cảm xúc.

Phân tích dữ liệu tô-pô (TDA) [4] cung cấp một giải pháp tiềm năng bằng cách trích xuất các đặc trưng hình học toàn cục, như persistence diagrams, từ bản đồ chú ý (attention maps) của PhoBERT. Các đặc trưng này cho phép phát hiện các mẫu ẩn, chẳng hạn như "lỗ hổng" tô-pô (topological holes) biểu thị sự không nhất quán ngữ nghĩa hoặc các cấu trúc liên kết phức tạp giữa các từ/câu, điều mà các mô hình học sâu thông thường không thể thực hiện. Kết hợp với kỹ thuật tăng cường dữ liệu (như thay thế từ đồng nghĩa, dịch ngược) [5], TDA có thể tăng cường khả năng tổng quát hóa của mô hình trên tập dữ liệu tiếng Việt như UIT-VSFC, đặc biệt cho các lớp thiểu số.

Đề tài đề xuất một phương pháp hybrid, tích hợp đặc trưng ngữ nghĩa từ PhoBERT, đặc trưng hình học từ TDA, và dữ liệu tăng cường để cải thiện hiệu suất phân lớp văn bản tiếng Việt. Phương pháp này không chỉ giải quyết các thách thức về

dữ liệu hạn chế và mất cân bằng lớp mà còn mở ra hướng nghiên cứu mới cho NLP tiếng Việt và các ngôn ngữ ít tài nguyên khác.

1.2 Các thách thức còn tồn tại

Phân lớp văn bản tiếng Việt đối mặt với các thách thức chính sau:

- *Dữ liệu hạn chế và không đồng đều:* Dữ liệu tiếng Việt được gắn nhãn chất lượng cao thường khan hiếm, đặc biệt cho các nhiệm vụ như phân tích cảm xúc hoặc phát hiện tin giả. Sự mất cân bằng giữa các lớp (ví dụ: ít mẫu cho cảm xúc tiêu cực so với tích cực) khiến mô hình dễ bị over-fitting hoặc thiên lệch (bias).
- *Hạn chế về biểu diễn toàn cục:* Các mô hình học sâu như PhoBERT chủ yếu tập trung vào ngữ nghĩa cục bộ và ngữ cảnh hai chiều, bỏ qua các mối quan hệ hình học toàn cục, như cấu trúc tô-pô của các từ đồng xuất hiện (co-occurrence) hoặc các chuỗi văn bản không liên tiếp. Điều này làm giảm khả năng phát hiện các mẫu phức tạp trong dữ liệu.
- *Đặc trưng ngôn ngữ tiếng Việt:* Tiếng Việt có cấu trúc từ ghép, thanh điệu, và ngữ cảnh văn hóa phức tạp, làm giảm hiệu quả của các kỹ thuật biểu diễn truyền thống như TF-IDF hoặc bag-of-words. Các mô hình học sâu cần được điều chỉnh để xử lý các đặc trưng này một cách hiệu quả.
- *Tích hợp TDA và tăng cường dữ liệu:* Các nghiên cứu gần đây [6] cho thấy TDA có thể trích xuất các mẫu tô-pô đặc trưng từ bản đồ attention, hữu ích cho việc phát hiện văn bản giả mạo. Tuy nhiên, việc tích hợp TDA với các mô hình tiếng Việt như PhoBERT chưa được nghiên cứu sâu. Đồng thời, các kỹ thuật tăng cường dữ liệu (như back-translation, synonym replacement) cần được điều chỉnh cẩn thận để tránh làm méo mó ngữ nghĩa, đặc biệt với tiếng Việt.

Sự kết hợp giữa TDA, kỹ thuật tăng cường dữ liệu, và PhoBERT vẫn là một khoảng trống nghiên cứu, đòi hỏi sự tối ưu hóa để cải thiện hiệu suất phân lớp trong các tập dữ liệu tiếng Việt không đồng đều.

1.3 Lý do chọn đề tài

Các mô hình phân lớp văn bản hiện tại, bao gồm cả PhoBERT, chủ yếu dựa trên phân tích ngữ nghĩa cục bộ hoặc theo chuỗi thời gian, chưa khai thác được các mối quan hệ hình học toàn cục trong dữ liệu văn bản. Ví dụ, trong bài toán phát hiện tin giả, các từ đồng xuất hiện toàn cục có thể tạo thành các "lỗ hổng" tô-pô (topological holes) biểu

thị sự không nhất quán về ngữ nghĩa, nhưng các mô hình như PhoBERT chưa tận dụng được thông tin này. Tương tự, trong phân tích cảm xúc, các mẫu tô-pô có thể giúp xác định các mối quan hệ phức tạp giữa các câu hoặc đoạn văn, nâng cao độ chính xác phân lớp.

Khởi phát từ các nghiên cứu tiên phong như [1] và [2], đề tài này tập trung vào việc biểu diễn văn bản tiếng Việt dưới dạng đồ thị tô-pô, kết hợp với kỹ thuật tăng cường dữ liệu để mở rộng tập huấn luyện và tích hợp với PhoBERT để cải thiện hiệu suất phân lớp. *Các lý do chính để chọn đề tài bao gồm*:

- *Tiếng Việt là ngôn ngữ ít tài nguyên:* Với dữ liệu gắn nhãn hạn chế, việc kết hợp tăng cường dữ liệu và TDA sẽ giúp cải thiện khả năng tổng quát hóa của mô hình, đặc biệt trong các ứng dụng thực tế như phân tích phản hồi sinh viên hoặc phát hiện tin giả.
- *Tiềm năng cải tiến hiệu suất:* Nghiên cứu của Kushnareva et al. [6] cho thấy tích hợp phân tích dữ liệu tô-pô (TDA) vào mô hình học sâu như BERT cải thiện độ chính xác từ 3–7% trong bài toán phát hiện văn bản giả mạo. Kết hợp TDA với PhoBERT, cùng với kỹ thuật tăng cường dữ liệu, được kỳ vọng nâng cao hiệu suất phân lớp thêm 5–10% so với PhoBERT cơ bản trên tập dữ liệu tiếng Việt như UIT-VSFC, nhờ khả năng khai thác cấu trúc tô-pô toàn cục và giảm thiểu over-fitting trên dữ liệu không đồng đều.
- *Ứng dụng thực tiễn cao:* Đề tài có tiềm năng ứng dụng trong các lĩnh vực như phân tích cảm xúc trên mạng xã hội, phát hiện tin giả, hoặc phân loại tài liệu y tế tại Việt Nam, nơi các giải pháp NLP hiệu quả còn hạn chế.
- *Đóng góp khoa học:* Nghiên cứu này không chỉ đóng góp vào lĩnh vực NLP tiếng Việt mà còn đặt nền tảng cho việc áp dụng TDA trong các ngôn ngữ ít tài nguyên khác, mở ra hướng nghiên cứu mới kết hợp giữa học sâu và phân tích hình học.

Đề tài sẽ sử dụng tập dữ liệu UIT-VSFC, một kho dữ liệu phản hồi sinh viên Việt Nam, để thử nghiệm và so sánh hiệu suất của mô hình đề xuất với các mô hình cơ bản như PhoBERT, LSTM, Bi-LSTM, và GRU, nhằm đánh giá mức độ cải tiến đạt được.

II. MỤC TIÊU, NỘI DUNG VÀ PHƯƠNG PHÁP

2.1 Mục tiêu của đề tài

Đề tài nhằm đạt được các mục tiêu chính sau, tập trung vào việc cải thiện hiệu suất phân lớp văn bản tiếng Việt thông qua sự kết hợp giữa kỹ thuật tăng cường dữ liệu,

phân tích dữ liệu tô-pô (Topological Data Analysis - TDA), và mô hình học sâu PhoBERT:

- *Tối ưu hóa hiệu suất phân lớp văn bản tiếng Việt:* Tích hợp các đặc trưng hình học từ TDA (như persistence diagrams và độ liên thông) vào biểu diễn ngữ nghĩa của PhoBERT [2], nhằm nâng cao khả năng nắm bắt các mối quan hệ toàn cục trong dữ liệu văn bản, đặc biệt trong các nhiệm vụ như phân tích cảm xúc và phân loại chủ đề.
- *Cải thiện xử lý dữ liệu hạn chế hoặc không đồng đều:* Kết hợp kỹ thuật tăng cường dữ liệu (data augmentation) với PhoBERT để mở rộng tập huấn luyện, giảm thiểu over-fitting và tăng khả năng tổng quát hóa của mô hình trong các tình huống dữ liệu tiếng Việt khan hiếm.
- *Xây dựng và đánh giá mô hình kết hợp:* Phát triển mô hình thực nghiệm trên tập dữ liệu tiếng Việt thực tế như UIT-VSFC (Vietnamese Students' Feedback Corpus), so sánh hiệu suất với các mô hình học sâu truyền thống (PhoBERT cơ bản, LSTM, Bi-LSTM, GRU), sử dụng các chỉ số như Accuracy, F1-score, Precision, Recall để chứng minh sự cải tiến.

Các mục tiêu này không chỉ giải quyết các hạn chế hiện tại trong NLP tiếng Việt mà còn cung cấp một khung nghiên cứu có thể mở rộng cho các ngôn ngữ ít tài nguyên khác.

2.2 Nội dung nghiên cứu

Nội dung nghiên cứu được xây dựng một cách hệ thống, bao quát từ cơ sở lý thuyết đến ứng dụng thực nghiệm, nhằm hỗ trợ việc đạt được các mục tiêu đề ra. Các nội dung chính bao gồm:

- *Cơ sở lý thuyết về PhoBERT, TDA và kỹ thuật tăng cường dữ liệu:* Tổng quan về kiến trúc PhoBERT (mô hình ngôn ngữ tiền huấn luyện cho tiếng Việt), các khái niệm cốt lõi của TDA (như persistence homology và diagrams), và các phương pháp tăng cường dữ liệu phổ biến trong NLP (Easy Data Augmentation - EDA, back-translation).
- *Tìm hiểu ưu điểm và nhược điểm của các phương pháp nghiên cứu gần đây:* Phân tích các mô hình hiện có liên quan đến phân lớp văn bản tiếng Việt, bao gồm ưu điểm của PhoBERT trong biểu diễn ngữ nghĩa và hạn chế của nó trong việc xử lý cấu

trúc hình học; đánh giá ứng dụng TDA trong NLP [3] và các nghiên cứu kết hợp học sâu với tăng cường dữ liệu.

- *Tiếp cận biểu diễn văn bản dưới dạng cấu trúc tô-pô:* Nghiên cứu cách chuyển đổi dữ liệu văn bản thành đồ thị tô-pô (topological graphs), kết hợp với tăng cường dữ liệu để mở rộng tập huấn luyện, đặc biệt cho các lớp dữ liệu thiểu số trong tập UIT-VSFC.
- *Phân tích các đặc tính tô-pô của dữ liệu văn bản thông qua PhoBERT:* Sử dụng bản đồ attention từ PhoBERT để trích xuất đặc trưng tô-pô, xác định các mẫu hình học như độ liên thông và chu trình bậc cao, nhằm bổ sung cho biểu diễn ngữ nghĩa.
- *Huấn luyện mô hình kết hợp TDA và PhoBERT để dự đoán phân lớp văn bản:* Triển khai mô hình hybrid, fine-tuning PhoBERT với đặc trưng tô-pô và dữ liệu tăng cường, đánh giá trên các nhiệm vụ phân lớp cảm xúc và chủ đề.
- *Nghiên cứu sự khác nhau của phương pháp đề xuất với các phương pháp khác:* So sánh hiệu suất mô hình đề xuất với các mô hình cơ bản, phân tích sự cải thiện nhờ TDA và tăng cường dữ liệu.
- *Phân tích kết quả đạt được và đề xuất hướng cải tiến sau này:* Đánh giá kết quả thực nghiệm, thảo luận các hạn chế, và gợi ý các hướng phát triển như mở rộng sang các tập dữ liệu khác hoặc tích hợp thêm các kỹ thuật TDA nâng cao.

2.3 Các phương pháp nghiên cứu

PhoBERT là một kiến trúc mạnh mẽ cho các bài toán NLP tiếng Việt, sử dụng tokenization cấp âm tiết và biểu diễn ngữ nghĩa dựa trên ngữ cảnh hai chiều. TDA sử dụng phép tương đồng liên tục (persistent homology) để trích xuất đặc trưng hình học toàn cục (như persistence diagrams), hiệu quả hơn các mô hình truyền thống trong việc kiểm tra đẳng cấu tô-pô và phát hiện các mẫu ẩn trong dữ liệu không gian cao chiều. Kỹ thuật tăng cường dữ liệu giúp mở rộng tập huấn luyện mà không cần thu thập thêm dữ liệu thực tế, đặc biệt hữu ích cho tiếng Việt – một ngôn ngữ ít tài nguyên.

Phương pháp nghiên cứu chính là kết hợp giữa biểu diễn ngữ nghĩa từ PhoBERT, đặc trưng hình học từ TDA, và tăng cường dữ liệu để tạo ra một mô hình hybrid nhằm cải thiện hiệu suất phân lớp văn bản. *Cụ thể, quy trình nghiên cứu được thực hiện theo các bước sau, với trọng tâm mở rộng vào cách triển khai kỹ thuật và tích hợp các thành phần*:

Bước 1: Biểu diễn văn bản thành đồ thị tô-pô: Dữ liệu văn bản từ tập UIT-VSFC được chuyển thành đồ thị (graph), nơi các đỉnh (nodes) đại diện cho từ hoặc câu, và các cạnh (edges) thể hiện mối quan hệ ngữ nghĩa hoặc tô-pô. Mối quan hệ tô-pô được xây dựng dựa trên bản đồ attention từ PhoBERT hoặc ma trận đồng xuất hiện từ các từ/câu trong văn bản.

Bước 2: Trích xuất đặc trưng tô-pô bằng TDA: Sử dụng TDA để phân tích đồ thị hoặc bản đồ attention từ PhoBERT, trích xuất các đặc trưng bền vững như persistence diagrams (biểu đồ bền vững).

Các đặc trưng này bao gồm độ liên thông (**dimension 0**: connected components), chu trình bậc cao (**dimension 1**: cycles/loops), giúp phát hiện các mẫu ẩn như sự không nhất quán trong văn bản giả mạo.

Bước 3: Tăng cường dữ liệu: Áp dụng các kỹ thuật EDA để mở rộng tập UIT-VSFC, đặc biệt cho các lớp thiểu số (ví dụ: cảm xúc tiêu cực). *Các phương pháp bao gồm*:

- Thay thế từ đồng nghĩa (synonym replacement) sử dụng từ điển tiếng Việt như VnCoreNLP.
- Chèn/xóa từ ngẫu nhiên (random insertion/deletion) với tỷ lệ 10-20% được chọn để cân bằng giữa việc tạo biến thể mới và bảo toàn ngữ nghĩa gốc của văn bản tiếng Việt, dựa trên thực nghiệm từ [5].
- Dịch ngược (back-translation) qua Google Translate (tiếng Việt -> tiếng Anh -> tiếng Việt) để tạo biến thể ngữ nghĩa.

Bước 4: Tích hợp và huấn luyện mô hình: Kết hợp đặc trưng tô-pô từ TDA (vector hóa từ persistence diagrams) với biểu diễn ngữ nghĩa từ PhoBERT (CLS token), đưa qua một tầng fully-connected với 256 nơ-ron, hàm kích hoạt ReLU, và dropout 0.3 để dự đoán nhãn phân lớp. Cơ chế chú ý (attention mechanism) trong PhoBERT được tinh chỉnh để ưu tiên các đặc trưng tô-pô, tăng cường khả năng phát hiện các mẫu toàn cục, từ đó cải thiện hiệu suất trên các lớp dữ liệu không đồng đều như lớp trung tính (~4.32%) trong UIT-VSFC.

Cơ chế điều hướng (attention mechanism) trong PhoBERT được tinh chỉnh để ưu tiên các đặc trưng tô-pô, giúp mô hình đạt hiệu suất cao hơn trong các lớp dữ liệu không đồng đều.

Kết hợp PhoBERT và TDA không chỉ dẫn đến cải thiện hiệu suất phân lớp mà còn cho phép phân loại các nút (nodes) dựa trên cấu trúc liên kết toàn cục. Các nghiên cứu [1, 2] chứng minh rằng cách tiếp cận này vượt trội hơn các phương pháp truyền thống, đặc biệt trong dữ liệu văn bản phức tạp như tiếng Việt.

III. TỔNG QUAN VỀ LĨNH VỰC NGHIÊN CỨU

3.1 Mô tả tập dữ liệu

Tập dữ liệu UIT-VSFC (Vietnamese Students' Feedback Corpus, version 1.0) [3] gồm hơn 16,000 câu phản hồi của sinh viên Việt Nam, được gắn nhãn thủ công cho hai nhiệm vụ:

- *Phân loại cảm xúc:* Nhãn gồm tiêu cực (0), trung tính (1), tích cực (2) (ví dụ: "Giảng viên nhiệt tình" – tích cực; "Cơ sở vật chất kém" – tiêu cực).
- *Phân loại chủ đề:* Nhãn gồm giảng viên (0), chương trình đào tạo (1), cơ sở vật chất (2), khác (3).

Đặc điểm nổi bật:

- *Độ đồng thuận annotator: >91% (cảm xúc), >71% (chủ đề).*
- *Hiệu suất cơ sở:* Mô hình Maximum Entropy đạt F1-score ~88% (cảm xúc), ~84% (chủ đề).
- *Chuẩn hóa dữ liệu:* Biểu tượng cảm xúc được chuyển thành văn bản (ví dụ: ":)" thành "colonsmile").
- *Tính phù hợp:* Số lượng mẫu lớn, hỗ trợ phân loại cảm xúc và chủ đề, phản ánh ngôn ngữ tự nhiên của sinh viên Việt Nam.
- *Phân bố lớp cảm xúc:* Tích cực ~49.69% (7,950 mẫu), tiêu cực ~45.99% (7,360 mẫu), trung tính ~4.32% (690 mẫu). Phân bố lớp chủ đề: Giảng viên ~36%, chương trình đào tạo ~27%, cơ sở vật chất ~20%, khác ~17%. Sự mất cân bằng, đặc biệt ở lớp trung tính và chủ đề "khác", đặt ra thách thức cho phân lớp chính xác.

Tập dữ liệu UIT-VSFC sẽ được sử dụng để huấn luyện và đánh giá mô hình, kết hợp tăng cường dữ liệu để mở rộng tập dữ liệu, đặc biệt cho các lớp thiểu số như cảm xúc tiêu cực hoặc chủ đề "khác"

3.2 Nghiên cứu trước đó

3.2.1 PhoBERT và các mô hình học sâu cho tiếng Việt

PhoBERT [2], được phát triển bởi Nguyen và cộng sự (2020), là một mô hình ngôn ngữ tiền huấn luyện dựa trên kiến trúc BERT, được tối ưu hóa cho tiếng Việt thông qua tokenization cấp âm tiết (syllable-based tokenization) sử dụng công cụ VnCoreNLP. PhoBERT tận dụng cơ chế attention hai chiều để nắm bắt ngữ cảnh, đạt hiệu suất vượt trội trong các nhiệm vụ NLP tiếng Việt, với F1-score cao hơn khoảng 5% so với BERT cơ bản trên các tập dữ liệu như VLSP (Vietnamese Language and Speech Processing). *Ví dụ*, trong phân lớp văn bản, PhoBERT đạt F1-score ~90% trên tập dữ liệu phân tích cảm xúc.

Tuy nhiên, PhoBERT chủ yếu tập trung vào biểu diễn ngữ nghĩa cục bộ và ngữ cảnh, chưa khai thác được các đặc trưng hình học toàn cục của dữ liệu văn bản, như các mối quan hệ tô-pô giữa các từ hoặc câu. Điều này khiến mô hình gặp khó khăn khi xử lý các tập dữ liệu không đồng đều hoặc khi cần phát hiện các mẫu phức tạp, chẳng hạn như sự không nhất quán trong văn bản giả mạo.

3.2.2 Phân tích dữ liệu tô-pô (TDA) trong NLP

Phân tích dữ liệu tô-pô (TDA) là một phương pháp toán học sử dụng phép tương đồng liên tục (persistent homology) để trích xuất các đặc trưng hình học bền vững từ dữ liệu, chẳng hạn như persistence diagrams biểu thị độ liên thông (dimension 0) và chu trình bậc cao (dimension 1). Trong NLP, Uchendu và Le [4] đã khảo sát ứng dụng TDA để phân tích bản đồ attention của các mô hình như BERT, phát hiện các mẫu tô-pô đặc trưng cho văn bản giả mạo. *Ví dụ*, các "lỗ hổng" tô-pô (topological holes) trong đồ thị từ bản đồ attention có thể chỉ ra sự không nhất quán ngữ nghĩa.

Tuy nhiên, ứng dụng TDA trong NLP tiếng Việt còn rất hạn chế, đặc biệt khi tích hợp với các mô hình như PhoBERT. Các nghiên cứu hiện tại chưa khám phá đầy đủ cách kết hợp đặc trưng tô-pô với biểu diễn ngữ nghĩa để cải thiện hiệu suất phân lớp văn bản tiếng Việt.

3.2.3 Kỹ thuật tăng cường dữ liệu trong NLP

Kỹ thuật tăng cường dữ liệu (Easy Data Augmentation - EDA) đã được áp dụng rộng rãi để mở rộng tập huấn luyện trong NLP, đặc biệt cho các ngôn ngữ ít tài nguyên như tiếng Việt [5]. *Các phương pháp phổ biến bao gồm*:

- *Thay thế từ đồng nghĩa (synonym replacement):* Thay thế các từ trong câu bằng từ đồng nghĩa, sử dụng từ điển tiếng Việt như VnCoreNLP.
- *Chèn/xóa từ ngẫu nhiên (random insertion/deletion):* Thêm hoặc xóa từ với tỷ lệ nhỏ (10-20%) để tạo biến thể.
- *Dịch ngược (back-translation):* Dịch văn bản sang ngôn ngữ khác (ví dụ: tiếng Anh) và dịch ngược lại để tạo ra các câu có ngữ nghĩa tương đương nhưng khác về cấu trúc.

Trong [3], Nguyen và cộng sự đã áp dụng EDA trên tập UIT-VSFC, cho thấy cải thiện hiệu suất khoảng 3-5% trong phân loại cảm xúc. Tuy nhiên, với tiếng Việt, các kỹ thuật này cần được điều chỉnh cẩn thận để tránh làm méo mó ngữ nghĩa, do đặc trưng từ ghép và thanh điệu.

3.2.4 Kết hợp TDA và học sâu

Trong [6], Kushnareva và cộng sự (2021) đã sử dụng phân tích dữ liệu tô-pô (TDA) để phân tích bản đồ chú ý của BERT, trích xuất các đặc trưng hình học như persistence diagrams để phát hiện văn bản giả mạo, đạt cải thiện độ chính xác từ 3–7% so với mô hình không dùng TDA. Ví dụ, các "lỗ hổng" tô-pô trong bản đồ chú ý giúp phát hiện sự không nhất quán ngữ nghĩa trong văn bản nhân tạo. Tuy nhiên, việc áp dụng TDA cho tiếng Việt còn hạn chế do đặc trưng ngôn ngữ phức tạp (như tokenization cấp âm tiết) và thiếu nghiên cứu tích hợp TDA với các mô hình như PhoBERT. Thách thức bao gồm việc điều chỉnh TDA để phù hợp với dữ liệu tiếng Việt và giảm độ phức tạp tính toán khi kết hợp với các mô hình học sâu.

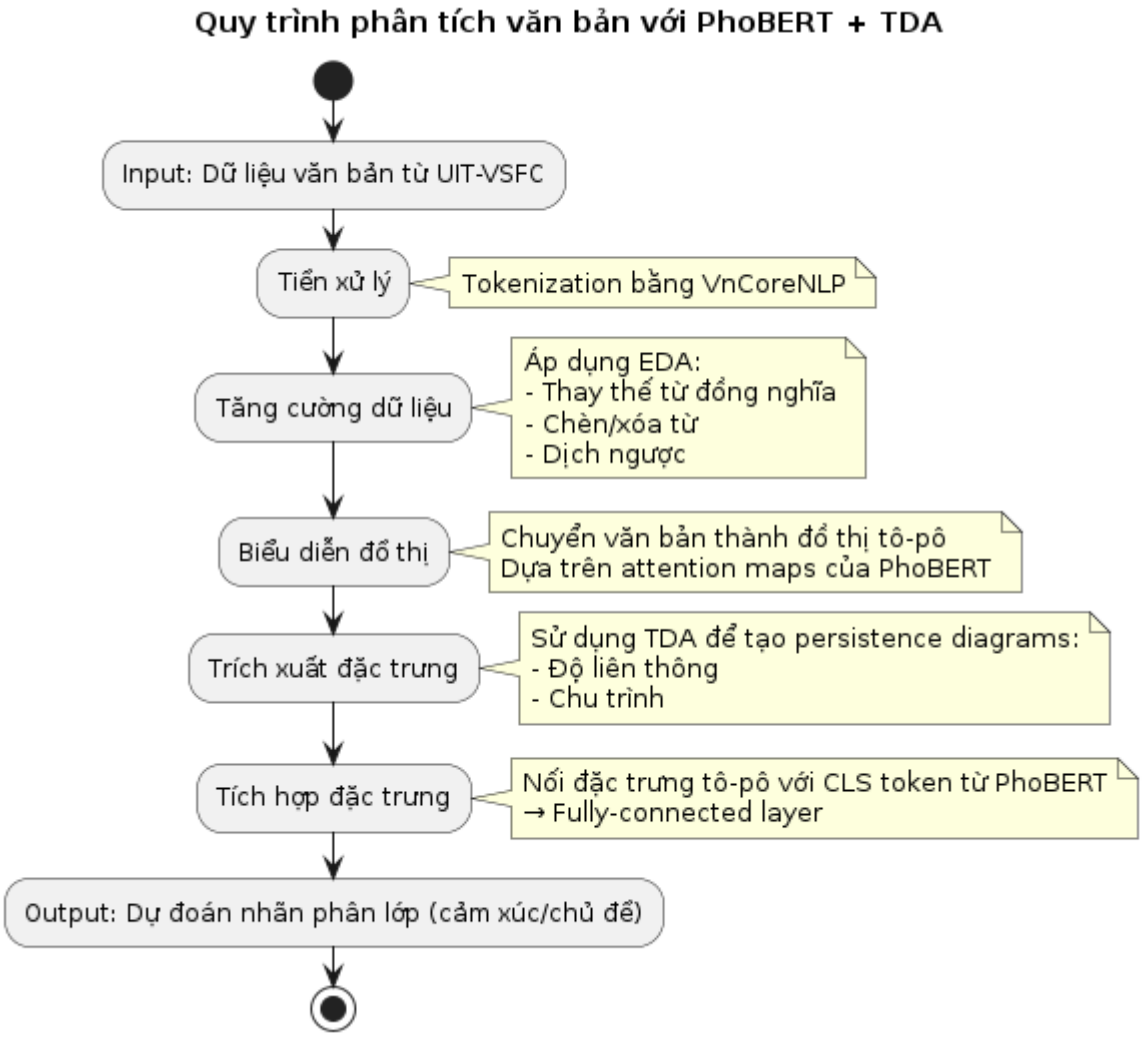
3.2.5 Các mô hình LSTM, Bi-LSTM, GRU

LSTM [1], Bi-LSTM, và GRU là các mô hình học sâu phổ biến trong phân lớp văn bản trước khi Transformer ra đời. LSTM sử dụng các cổng (forget, input, output) để giải quyết vấn đề vanishing gradient, Bi-LSTM cải thiện bằng cách xử lý dữ liệu hai chiều, và GRU đơn giản hóa với cổng update và reset, giảm chi phí tính toán. Các mô hình này được chọn làm baseline vì chúng từng là tiêu chuẩn cho các nhiệm vụ NLP như phân tích cảm xúc, nhưng bị hạn chế bởi việc chỉ xử lý dữ liệu chuỗi, không khai

thác được cấu trúc tô-pô toàn cục của dữ liệu văn bản phức tạp như UIT-VSFC. Do đó, chúng phù hợp để so sánh với phương pháp Transformer kết hợp TDA đề xuất trong nghiên cứu này.

3.3 Thiết lập và xây dựng mô hình

Quy trình xây dựng mô hình bao gồm các bước chính sau, nhằm tích hợp mô hình PhoBERT, phân tích dữ liệu tô-pô (TDA), và kỹ thuật tăng cường dữ liệu để cải thiện hiệu suất phân lớp văn bản tiếng Việt trên tập dữ liệu UIT-VSFC:



Hình 1:Minh họa quy trình tích hợp PhoBERT, TDA, và tăng cường dữ liệu

- *Biểu diễn văn bản thành đồ thị tô-pô:* Văn bản từ tập UIT-VSFC được chuyển thành đồ thị, trong đó các đỉnh (nodes) đại diện cho từ hoặc câu, và các cạnh (edges) thể hiện mối quan hệ ngữ nghĩa (dựa trên bản đồ chú ý của PhoBERT) hoặc tô-pô (dựa trên ma trận đồng xuất hiện từ).

- *Phân tích tô-pô bằng TDA:* Sử dụng TDA để trích xuất các đặc trưng hình học từ bản đồ chú ý (attention maps) của PhoBERT, bao gồm persistence diagrams biểu thị độ liên thông (dimension 0) và chu trình bậc cao (dimension 1). Các đặc trưng này được chuyển thành vector đặc trưng thông qua vector hóa persistence diagrams (ví dụ: sử dụng persistence images).

- *Tăng cường dữ liệu:* Áp dụng các kỹ thuật như thay thế từ đồng nghĩa (dựa trên VnCoreNLP), chèn/xóa từ ngẫu nhiên (tỷ lệ 10–20%), và dịch ngược (back-translation) để mở rộng tập dữ liệu, đặc biệt cho các lớp thiểu số như cảm xúc tiêu cực hoặc chủ đề "khác".

- *Tích hợp và huấn luyện mô hình:*

- Cấu hình PhoBERT: Sử dụng PhoBERT-base với 12 tầng Transformer, 768 chiều ẩn, 12 đầu chú ý, và khoảng 135 triệu tham số. Mô hình được tinh chỉnh (fine-tuned) trên tập UIT-VSFC, sử dụng 3 tầng cuối cùng với learning rate 2e-5 và optimizer AdamW để tối ưu hóa hiệu suất phân lớp.

- Tích hợp đặc trưng tô-pô: Vector đặc trưng tô-pô (từ persistence diagrams) được nối (concatenated) với đầu ra ngữ nghĩa từ tầng cuối của PhoBERT (CLS token). Kết quả được đưa qua một tầng fully-connected với 256 nơ-ron, hàm kích hoạt ReLU, và dropout 0.3 để dự đoán nhãn phân lớp (cảm xúc hoặc chủ đề).

- Huấn luyện: Mô hình được huấn luyện trên GPU (ví dụ: NVIDIA V100), với batch size 16, trong 5–10 epoch, sử dụng hàm mất mát cross-entropy.

- Cấu hình này đảm bảo kết hợp hiệu quả giữa biểu diễn ngữ nghĩa cục bộ từ PhoBERT và đặc trưng hình học toàn cục từ TDA, đồng thời giảm thiểu over-fitting thông qua tăng cường dữ liệu, từ đó nâng cao hiệu suất phân lớp văn bản tiếng Việt.

3.4 Đánh giá

Hiệu suất của mô hình được đánh giá trên tập dữ liệu UIT-VSFC, sử dụng các chỉ số chuẩn trong NLP:

- *Accuracy:* Tỷ lệ dự đoán đúng trên tổng số mẫu.

- *F1-score:* Trung bình điều hòa giữa Precision và Recall, đặc biệt hữu ích cho các tập dữ liệu không cân bằng.

- *Precision:* Tỷ lệ dự đoán đúng trong số các dự đoán tích cực.

- *Recall:* Tỷ lệ các mẫu tích cực được dự đoán đúng.

- Các độ đo này được chọn vì phù hợp với bài toán phân lớp văn bản tiếng Việt: F1-score ưu tiên cho dữ liệu không cân bằng (như lớp trung tính chiếm ~4.32% trong UIT-VSFC), Accuracy đánh giá hiệu suất tổng quát, còn Precision và Recall cung cấp phân tích chi tiết về hiệu quả trên từng lớp cảm xúc và chủ đề.

3.5 So sánh với các phương pháp khác

Mô hình đề xuất (PhoBERT + TDA + tăng cường dữ liệu) sẽ được so sánh với các mô hình cơ bản trên tập dữ liệu UIT-VSFC:

- *PhoBERT cơ bản:* Mô hình PhoBERT không tích hợp TDA hoặc tăng cường dữ liệu, sử dụng trực tiếp từ thư viện Hugging Face.
- *LSTM:* Mô hình xử lý chuỗi thời gian, sử dụng các cổng để kiểm soát luồng thông tin.
- *Bi-LSTM:* Phiên bản hai chiều của LSTM, cải thiện khả năng nắm bắt ngữ cảnh.
- *GRU:* Mô hình đơn giản hơn LSTM, với hai cổng update và reset.

Thực nghiệm sẽ được tiến hành với cùng điều kiện: tỷ lệ chia tập dữ liệu theo chuẩn UIT-VSFC (70% train: 11,426 mẫu, 10% dev: 1,538 mẫu, 20% test: 3,166 mẫu), sử dụng GPU NVIDIA V100 hoặc tương đương, framework PyTorch với batch size 16 và learning rate 2e-5. Các chỉ số đánh giá bao gồm Accuracy, F1-score, Precision, và Recall, được tính cho cả hai nhiệm vụ phân lớp cảm xúc và chủ đề. Dự kiến, mô hình đề xuất sẽ vượt trội nhờ khả năng kết hợp đặc trưng tô-pô (phát hiện các mẫu toàn cục) và dữ liệu tăng cường (giảm over-fitting). Nghiên cứu của Kushnareva et al. [6] cho thấy TDA cải thiện độ chính xác 3–7% trong phát hiện văn bản giả mạo, và đề tài này kỳ vọng đạt cải tiến tương tự trong phân lớp văn bản tiếng Việt.

IV. TIẾN ĐỘ THỰC HIỆN ĐỀ TÀI

4.1 Các bước thực hiện, thời gian bắt đầu và hoàn thành

TT	Tháng	4	5	6	7	8	9	10
1	Chuẩn bị thực hiện đề cương luận văn							
2	Thu thập tài liệu, nghiên cứu cơ sở lý thuyết về PhoBERT, TDA và tăng cường dữ liệu							
3	Xây dựng và tiền xử lý tập dữ liệu UIT-VSFC, áp dụng kỹ thuật tăng cường dữ liệu							
4	Triển khai mô hình kết hợp PhoBERT với TDA, trích xuất đặc trưng topo từ attention maps.							

5	Huấn luyện, thử nghiệm mô hình trên tập dữ liệu và đánh giá hiệu suất (Accuracy, F1-score, v.v.)							
6	So sánh kết quả với các mô hình cơ bản, phân tích và đề xuất cải tiến							
7	Viết luận văn, chỉnh sửa và hoàn thiện đề tài							

Bảng 1: Tiến độ thực hiện đề tài

4.2 Bố cục dự kiến của luận văn

Dự kiến đề tài sẽ bao gồm 5 mục chính:

CHƯƠNG 1: GIỚI THIỆU

1.1 Đặt vấn đề

1.2 Các thách thức còn tồn tại

1.3 Lý do chọn đề tài

CHƯƠNG 2: TỔNG QUAN VỀ LĨNH VỰC NGHIÊN CỨU

2.1 Nghiên cứu trước đó

2.2 Mô tả tập dữ liệu (UIT-VSFC)

2.3 Thiết lập và xây dựng mô hình

CHƯƠNG 3: PHƯƠNG PHÁP LUẬN

3.1 Phương pháp nghiên cứu

3.2 Trích xuất đặc trưng tô-pô (TDA)

CHƯƠNG 4: THỰC NGHIỆM VÀ ĐÁNH GIÁ

4.1 Thiết lập thử nghiệm và cấu hình mô hình

4.2 Các thang đánh giá (Accuracy, F1-score, Precision, Recall)

4.3 Cơ sở so sánh với các mô hình khác

4.4 Kết quả thực nghiệm, phân tích lỗi và đánh giá

CHƯƠNG 5: KẾT LUẬN & HƯỚNG PHÁT TRIỂN

5.1 Khó khăn và thuận lợi của đề tài

5.2 Kết quả thực hiện của đề tài so với dự định ban đầu

5.3 Kết luận, đánh giá và hướng phát triển

V. TÀI LIỆU THAM KHẢO

[1] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural Comput., vol. 9, no. 8, pp. 1735–1780, 1997.

[2] D. Q. Nguyen and A. T. Nguyen, “PhoBERT: Pre-trained language models for Vietnamese,” in Findings of the Association for Computational Linguistics: EMNLP 2020, 2020, pp. 1037–1042.

[3] K. V. Nguyen et al., “UIT-VSFC: Vietnamese Students’ Feedback Corpus for Sentiment Analysis,” in Proc. 10th Int. Conf. Knowledge Syst. Eng. (KSE), 2018, pp. 19–24.

[4] A. Uchendu and T. Le, “Unveiling topological structures in text: A comprehensive survey of topological data analysis applications in NLP,” arXiv preprint arXiv:2411.10298, 2024.

[5] J. Wei and K. Zou, “EDA: Easy data augmentation techniques for boosting performance on text classification tasks,” in Findings of the Association for Computational Linguistics: EMNLP 2019, 2019, pp. 6382–6388.

[6] L. Kushnareva et al., “Artificial text detection via examining the topology of attention maps,” in Findings of the Association for Computational Linguistics: EMNLP 2021, 2021, pp. 635–649.